Framework for Transitioning from IASA to NCLB

Series Overview

September 2004

by
Doris Redfield, Ph.D. and
Jan Sheinker, Ed.D.

with
CAS SCASS Study Group: Transitions in Assessment from IASA to NCLB

Acknowledgments

This paper resulted from the work of the Study Group on Transitions in Assessments from IASA to NCLB comprised of state educational specialists and consultants of the Comprehensive Assessment Systems for ESEA Title I (CAS) State Collaborative on Assessment and Student Standards (SCASS). The members of the Study Group benefited from discussions among SCASS colleagues throughout 2003 and 2004:

Robert Anderson, California (former Chair)
Mildred Bazemore, North Carolina
Dale Carlson, Consultant
Tim Crockett, Measured Progress
Carol Crothers, Nevada (Chair)
William J. Erpenbach, Consultant
Arthur Halbrook, CCSSO
Ellen Hedlund, Rhode Island
Tammy Howard, North Carolina
Susan Ketchum, Wisconsin
Bernadette Morris, Louisiana
Les Morse, Alaska
Jason Nicholas, Wyoming
Grace Ross, ED, Ex-Officio
Alan Sheinker, CTB
Rodney Watson, Louisiana
Charles Wayne, Pennsylvania

Jan Sheinker, CAS SCASS Coordinator

This paper was supported entirely by funding from member States of the Comprehensive Assessment Systems for ESEA Title I State Collaborative on Assessment and Student Standards (CAS SCASS), through the Council of Chief State School Officers. Information about the CAS SCASS is available on the CCSSO website, http://www.ccsso.org.

This publication and any comments, observations, recommendations, or conclusions contained herein reflect the work of the authors. They do not necessarily reflect the views of the Council of Chief State School Officers.

The *Improving America's Schools Act of 1994* (IASA), implemented in 1995, was the reauthorization of the *Elementary and Secondary Education Act* of 1965 (ESEA). To receive Title I funds, IASA, required states to establish challenging academic content and student performance standards for *all* students. IASA also required states to assess all students relative to these standards and to demonstrate that students were making "adequate" progress toward achieving the standards. These assessments were to take place at least once annually within each of three grade spans: 3-5, 6-9, and 10-12, and to cover reading or language arts and mathematics. The ultimate goal was for the United States to be first in the world in student achievement.

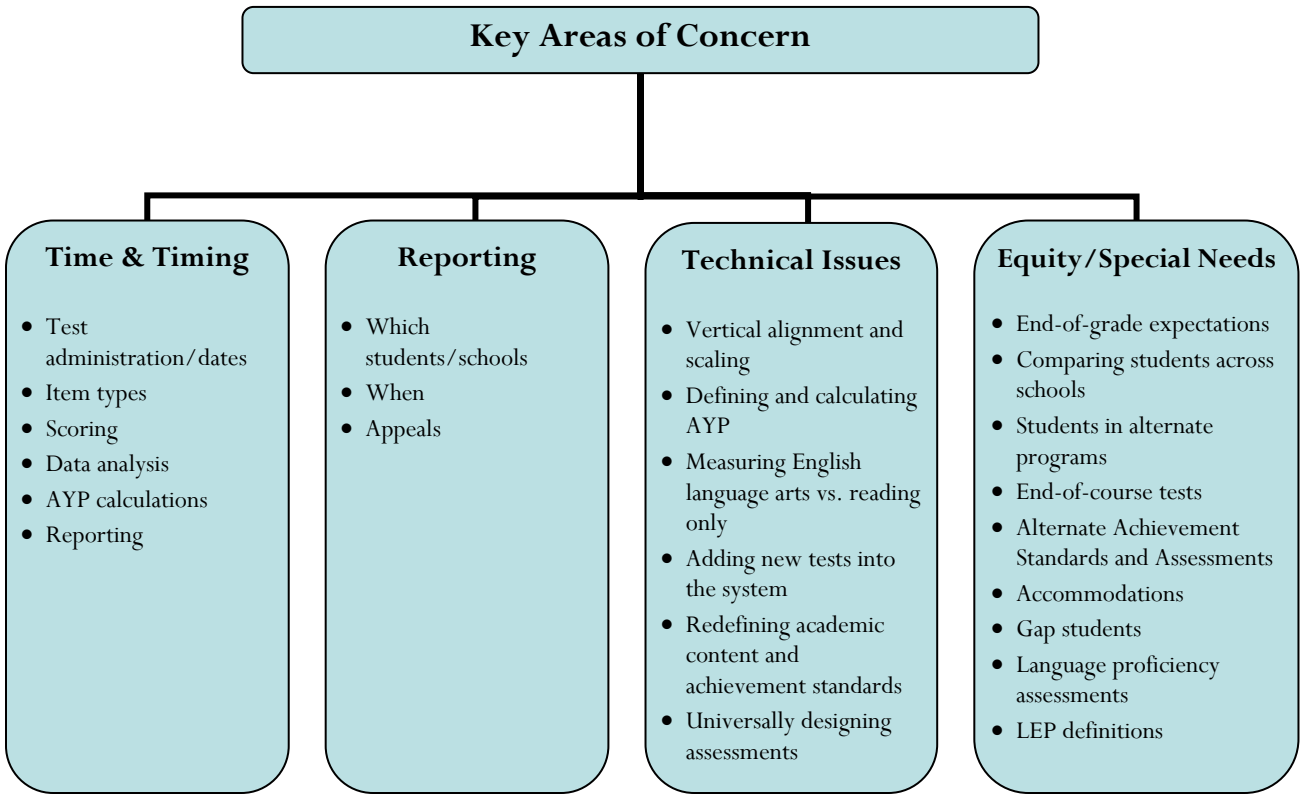
The 2001 reauthorization of ESEA, the *No Child Left Behind Act of 2001* (NCLB), placed additional challenges on states that were already struggling to meet IASA requirements. NCLB established deadlines for states to expand the scope and frequency of student testing (annually in every grade 3 through 8 and once in grade 10 through 12); to revamp their accountability systems; to ensure that every student is taught by a qualified teacher; to demonstrate "adequate yearly progress" regarding the percentages of students achieving proficiency in reading or language arts and mathematics; to close the achievement gap between disadvantaged and advantaged students; and to base practices on findings from scientific research. The ultimate goal is proficiency for every student relative to a set of standards.

IASA emphasized standards and assessments. NCLB emphasizes accountability, i.e., holding states, local education agencies (LEAs), and schools accountable for student performance on assessments that are aligned with rigorous academic content and achievement standards (designated as performance standards by IASA). Furthermore, NCLB requirements became effective immediately, whereas IASA requirements included time for transitioning from existing systems to new requirements. The accountability requirements of NCLB and the associated high-stakes consequences of failure to make adequate yearly progress (AYP) can stress the assessment systems that provide the core data for determining AYP status. Further, requirements to disaggregate assessment data by specified student populations may result in interpretations of questionable quality, particularly when the numbers of students in specific groups may be small.

In their 2002 publication, ***Making Valid and Reliable Decisions in Determining Adequate Yearly Progress***, Marion, White, Carlson, Erpenbach, Rabinowitz, and Sheinker provide a useful summary of the key differences between the accountability requirements of IASA and NCLB (see Appendix A). While those authors chiefly address accountability systems under NCLB, the purposes of this paper—the first in a series of issue papers—are (1) to identify implementation issues surrounding the transition of assessments from IASA to NCLB, (2) to provide a framework for thinking about these assessment issues, (3) to provide a tool to guide decisions related to the implementation of NCLB assessment requirements, and (4) to foreshadow issues to be addressed in subsequent papers. This paper is primarily for policy "influencers" such as state education agency leaders, particularly directors of assessment and Title I programs, who interact with and influence policymakers and other key decision makers.

Framework

Before discussing the implementation issues related to transitioning from IASA to NCLB, it might be useful to place them in a framework (see below). This framework sorts key areas of concern by issue categories: time and timing (e.g., deadlines and implications for scheduling); reporting (e.g., who gets assessed and what scores are reported to whom); technical issues, such as alignment and the measurement of AYP; and issues of equity, including meeting the needs of special populations (e.g., students with disabilities, English language learners).



Issues

In most cases, the issues presented in this paper have no single or easy answers. For that reason, it is important for states to consider them and to have clear and defensible rationales for why they address a particular issue in a particular way. Otherwise, the validity of the state's assessment system and, therefore, the accountability decisions based on that system will be eroded.

Time and Timing

Time and timing cannot be separated from sequencing. For example, if grade-level expectations (GLEs) and performance descriptors have not been developed, it is difficult to develop assessments that align with those GLEs and descriptors. Alignment is critical to the validity of specific assessments as well as to the validity of overall assessment and accountability systems. NCLB requires an aligned system. NCLB states that student academic achievement standards must be developed using a formal process, must be grade specific, and must be formally approved by the entity responsible for standards policies.

When thinking about time, states need to consider both annual requirements and within-year requirements. The NCLB annual requirements are summarized in Appendix B, which indicates when the various elements of the state assessment must be in place. These elements, such as implementing reading or language arts and mathematics assessments to all students in grades 3 through 8 by 2005-06 are fairly straightforward.

Perhaps more challenging are the within-year requirements shown in Appendix C. For example, Section 1116(b)(2)(C) of the law states that after an LEA) provides a school with the opportunity to review the data used to identify the school as "in need of improvement," the LEA must make a final and public determination of the school's status within 30 days of providing the opportunity for review.

Section 1116(b)(3)(E) states that the LEA, within 45 days of receiving a school plan, shall (1) establish a peer review process to assist with review of the plan; (2) promptly review the school plan, working with the school as necessary; and (3) approve the plan if it meets the requirements for approval.

Section 1116(c)(5)(B) of the law asserts that if a LEA believes that its identification by the State education agency (SEA) is in error, for statistical or other substantive reasons, the LEA may provide supporting evidence to the state education agency (SEA), which shall consider the evidence before making a final determination. This determination must be made not later than 30 days after the SEA provides the LEA with the opportunity to review the data.

However, Section 1116(c)(10)(D) states that prior to implementing any corrective action, the SEA shall provide notice and a hearing to the affected local agency, if state law provides for such notice and hearing. The hearing shall take place not later than 45 days following the decision to implement corrective action.

Taken together, these sections of the law present serious timing dilemmas. Within 30 days after identification, districts have 30 days to challenge the identification, but the district has only a 45-day window in which to challenge the state. This suggests that local and state windows for challenge may need to occur simultaneously in the interest of time. Otherwise, LEAs would only have 15 days in which to accomplish their appeals process.

Test administration. Timing the administration of tests is critical because NCLB requires AYP to be calculated from the time the data are received. Clearly, it is advantageous to

allow as much time to lapse as possible (a "full academic year" or FAY) between calculations so as much instruction as possible occurs between calculations. The timing of test administration becomes particularly problematic when performance assessments that cannot be machine scored are administered. Erpenbach, Forte Fast, and Potts (2003) provide numerous helpful illustrations of how states are handling this issue. Approaches range from changing the assessment window from late winter to late spring or fall, to defining a FAY as continuous enrollment from one test administration to the next or as "no more than 365 days."

Scoring. A number of states (e.g., Idaho, Maine, Oregon, Utah, Virginia) use or are moving toward online administrations of tests to allow for more rapid scoring and reporting of results. Idaho administers its state assessment online and incorporates a common set of grade-level, standards-aligned items to counterbalance the adaptive nature of the test. However, the degree of alignment remains a challenge for the state. The U.S. Department of Education (ED) has consistently raised alignment as an issue on adaptive tests and has not fully approved such a model to date. Maine will pilot an eighth-grade writing assessment in spring 2004 and hopes to administer the full assessment to eighth graders online by next year, depending on findings from the pilot test. Oregon offers its assessments in both online and paper-pencil versions. However, a number of states, including Georgia and South Dakota, have abandoned plans for online state assessments due to access and alignment challenges. Access remains a major barrier to the widespread use of online assessment due to both limited availability of computers and to difficulty in defining participation so that reliable and consistent data can be collected within a defined testing window, across diverse testing sites, and with varying equipment. All these factors can influence the reliability and comparability of assessment results collected in this manner.

Item types. While online assessment can be a useful strategy for machine-scorable tests and items, they may negatively impact states' willingness to include assessments that are performance-based and/or cannot be machine scored. Theoretically, online assessment should offer the advantage of being able to present simulations and new item types that offer cognitive challenge but can be computerized and scaled. However, technical challenges and cost remain barriers to their inclusion in state tests. Although the issue for states is turn around time of test results for making AYP decisions, choices to exclude constructed-response and performance-based items may have implications for the degree of alignment of assessments with academic content standards. Both IASA and NCLB call for full alignment of assessments with the range, breadth, and depth; level and degree of cognitive complexity; and comprehensiveness of the standards. Aligning assessments to the level of difficulty of the concepts and processes described in the standards must be clearly reflected in the assessment plan, blueprint, and items and tasks. While the compromise of alignment can be an unintended consequence of online testing, it may not be the only one. Excluding constructed-response items also may impact instruction if teachers perceive that the state and district place a priority on what is on the test.

Data analysis. Data analysis is closely linked to reporting requirements. It is important to analyze data according to the required disaggregate categories (race/ethnicity, economically disadvantaged, LEP, students with disabilities). The need to correctly categorize students by their appropriate disaggregate group membership has implications for database accuracy. Many states are working to create databases that include individual student data for the first time in the belief

that only through this mechanism will it be possible to conduct the required data analyses with any degree of accuracy. Further, the requirement to calculate AYP on receipt of the data means that data need to be analyzed expeditiously. And, because high-stakes decisions will be based on the results of the analyses, the accuracy of the data as well as the error surrounding scores become critical considerations. In fact, the law requires that assessments be valid and reliable, and that results be reported in a manner to ensure the accurate interpretation of results. ED's *Information Quality Guidelines* (2002) further emphasizes the importance of reporting data in a manner that confirms and documents the reliability of the data and acknowledges shortcomings and error in the results. The range within which a score can be described as accurate is important to the reliable interpretation of the results. Reporting measurement-error range or confidence intervals, reveals the range within which scores can be interpreted accurately and clarifies the meaning of scores.

Adequate Yearly Progress (AYP) calculations. Procedures for calculating AYP are beyond the scope of this paper. For a comprehensive treatment of AYP, readers are referred to Carlson (1996); Erpenbach, Forte Fast & Potts (2003); Marion, White, Carlson, Erpenbach, Rabinowitz, and Sheinker (2002); and Winter (1996). In most cases, AYP is based on a specified percentage of students reaching a standard of achievement. AYP must be calculated on receipt of the assessment data, and AYP information is necessary for notifying schools and LEAs of their identification status in accordance with the timeline requirements of NCLB as described earlier. The implication for absolute accuracy in assessment data used to calculate AYP is evident in the severity of the sanctions for failure to meet AYP required by the NCLB Act.

Reporting. While reporting is an issue in its own right, it is also an issue of time and timing. States are charged with making AYP-based identifications before the beginning of each school year so parents can be notified of their right to school choice and supplemental educational services under the provisions of NCLB. This creates tension between the time requirements for reporting and the time requirements for FAY. One approach to dealing with this issue is to base determinations on preliminary data. This is the approach being taken by many states (e.g., California, Idaho, Kentucky, Louisiana, Massachusetts, Mississippi, Missouri, Montana, New Jersey, Oklahoma, Pennsylvania, South Carolina, Utah, West Virginia, Wisconsin, and Wyoming). With this approach, mechanisms must be in place for ensuring that final actions are based only on final and accurate findings. In the interest of timely reporting during this process, states are encouraged to report results as data become available rather than waiting for all elements of a report card to become available.

Reporting Issues

A comprehensive treatment of NCLB reporting requirements and associated issues is provided by Erpenbach, Forte Fast, and Potts (2003). The following discussions highlight some of those issues.

Which Schools? The law is clear that all schools are to be held accountable. When schools or districts are too small to report on subgroups, they must base AYP decisions on the school or district as a whole. The definition of "too small" varies from state to state. Thirty-three

states require a minimum "n" size of 10 students for reporting, although several states require as few as 5 and a few states require as many as 20. Fourteen states require a minimum "n" size of 30 while a dozen more require 40 in a cell for making AYP decisions. However, the number can be as small as the number required to meet a test of statistical significance in North Dakota or as many as 100 in California or 200 in Texas under certain conditions. The extremely small rural schools of states like Idaho, North and South Dakota, and Wyoming present unique challenges for both reporting and accountability.

New schools are an additional concern for reporting. In the case of new schools, AYP determinations do not need to be made until the end of the second school year following the opening of the school. Schools created as a result of reorganizations, such as merging or rerouting intact student populations or changing status to a charter school, do not qualify as new schools.

In addition to questioning which schools' scores and status are reported and how, it is important to wrestle with the issue of individual student scores. For reporting purposes, student identity must be protected, but for diagnostic purposes student identity is critical. NCLB requires that diagnostic information be reported at the individual level, one more reason many states are moving toward establishing databases with data that are identifiable at the individual student level. Reporting individual student scores also has necessitated changes in some state assessments that, under IASA, used methodologies that permitted full alignment of the assessment with the standards at the school and district level, but sampled the standards at the individual student level. The NCLB requirement that each student's scores be reported at the standards level for diagnostic purposes places a greater burden on the individual student score and on the comprehensiveness with which each student is assessed. Assessing each standard adequately to make meaningful judgments at the individual student level could lengthen tests significantly.

What is reported? NCLB also requires "item analysis," although this is interpreted as being at the strand level rather than at the individual item level. There is no expectation to release individual items. The intent is to provide information at the standard level for diagnostic purposes to parents, teachers, and the public. Confusion arises, however, in the precise meaning of the terms "strand" and "standard" for each state. Each state organizes its standards differently from other states and may even organize them differently from one content area to another. For example, a state may refer to reading, writing, speaking, and listening as the "strands" within language arts; and to numbers and operations, algebra, geometry, measurement, and data analysis as the "strands" within mathematics. Standards may be identified by a state as global statements for each of these strands or as more specific content requirements related to these global statements. Grade-level expectations or grade-level standards may be further subsets of these. To meet the "item analysis" and "standards reporting" requirements of NCLB, some states cluster the statements they identify as standards into headings such as information text, literary text, and research, or into basic reading, literal comprehension, and inferential comprehension to report reading scores for accountability purposes. The extent to which this information is interpretable for individual students even for diagnostic purposes is, however, a matter of concern to states.

When? Should final identifications of schools in need of improvement be made before or after the time permitted for appeals? If the appeals process described in NCLB is followed, final identifications would be made after the appeals process; however, tight timelines prescribed by the NCLB Act have caused some states to make determinations on the basis of preliminary (pre- appeals) designations and later modify the list as necessary, based on the results of quality assurance procedures and appeals. For example, if a school in Wisconsin is identified as needing improvement, the identification is labeled as "preliminary," and the school has 30 days in which to appeal before a final designation is made. South Carolina and Massachusetts also issue preliminary reports to speed up provision of choice and supplemental services. In practice, some districts have resisted complying until final determinations are made. Short timelines for reporting also increase pressure on the assessment system by shortening the window for data auditing that is necessary for quality checks on scoring procedures and on coding for accurate disaggregation of data.

Appeals. The appeals requirements are spelled out in the "Timing" section of this document. Remaining, however, is the question of the state's role in monitoring local processes. How can the monitoring be feasibly and reliably accomplished? Wisconsin provides an example of how appeals may be handled at the state level. In this case, the state designs reports to be sent home to families. The appeals template is on the Wisconsin Department of Public Instruction's website (http://www.dpi.state.wi.us/oea/doc/reconsid02.doc).

Technical Issues

Vertical alignment and scaling. A vertical scale links (grade) levels of a test in order for scores to be compared from one grade to the next. To illustrate, a fourth-grade student would take an assessment that includes items from the third- and fifth-grade tests in order to create a vertical scale. Vertical scaling is thought by some to be one solution to the fundamental problem of longitudinally interpreting assessment results (e.g., from year-to-year or grade-to-grade). While NCLB requires comparing test results from year-to-year, it is not clear what kind of scaling is needed to make such comparisons. For example, are traditional vertical scaling techniques valid for purposes of calculating AYP, and/or do these techniques add value to the process of determining AYP based on test results? It is uncertain whether the trait in question is unidimensional and consistent enough over grade levels to make the use of vertical scales practical in this application. Such application is particularly questionable within the context of standards-based testing. If standards are vertically aligned so there is little redundancy from one grade to the next, vertical scaling may, in fact, be undesirable.

For these and other reasons, vertical scaling remains controversial. While the idea of vertical scaling may appear attractive on the surface—because it can create the illusion of score comparability across grade levels and of a particular difference between scores representing a year's growth—judgments about students' proficiency on grade-level standards may be confounded by the inclusion of items based on standards from other grade levels. While the idea of vertical scaling may be statistically interesting, it may also be misleading or misused.

For example, what if the test for one grade level contains items that are more difficult than the items at another grade level, but those items align with the grade-level standards being assessed? It could be that the standards require revision to be vertically aligned from grade to grade in terms of depth of knowledge, range of knowledge, balanced representation, and source of challenge (see Webb, 2002, for a comprehensive treatment of alignment). States have found it challenging to differentiate standards from one grade span to another; clearly differentiating standards from one grade level to another is even more challenging. And, vertical alignment seems an important pre-requisite to any meaningful application of vertical scaling techniques.

Linking of horizontal scales may be less difficult and just as valuable as vertical scaling in the interpretation of results. The topic of vertical scaling continues to be the subject of much discussion and research, particularly in light of the movement toward developing grade-level expectations (GLE). For example, the Technical Issues in Large-Scale Assessment State Collaborative on State Standards and Assessments (TILSA SCASS) is currently working with Laurie Wise to address the applications of vertical alignment to NCLB testing. Robert Lissitz's and Huynh Huynh's (2003) discussions of vertically moderated standards may also shed light on these questions.

Defining and calculating Adequate Yearly Progress (AYP). The formula for calculating AYP is an ongoing issue. The Joint Study Group of the Accountability Systems and Reporting (ASR) and the Comprehensive Assessment Systems for ESEA Title I (CAS) State Collaboratives on Assessment and Student Standards (SCASS) are doing work in this area. For an in-depth treatment of this topic, see the ASR-CAS SCASS Joint Study Group publication, ***Making Valid and Reliable Decisions in Determining Adequate Yearly Progress*** (Marion, White, Carlson, Erpenbach, Rabinowitz, & Sheinker, 2002).

Measuring English language arts (ELA) versus reading only. A number of states are dropping their writing assessments and using only tests of reading to assess ELA. Major reasons include the time it requires to score writing assessments and the expense of doing so. There is also debate about how to treat reading and writing scores in the calculation of AYP. Should they be treated separately or combined? If separately, which score *really* counts? ED has not been clear in its guidance relative to this issue. However, states approved for ELA standards and assessments under IASA that have elected to assess only reading under NCLB are likely to be subject to re-review of their standards and assessments. Erpenbach, Forte Fast, and Potts (2002) provide a number of examples of state approaches to the issue, ranging from weighting reading more heavily than writing (e.g., Delaware) to including writing assessments only in certain years (e.g., Florida).

Several states that included writing assessments in their NCLB plans are grappling with the question of whether to continue these when they implement grades 3 through 8 testing. The added expense of testing writing at every grade level may be too much for fiscally challenged states. It is unclear whether ED would approve a system in which writing continues to be part of the grade-span assessments rather than grade-level assessments if they are to part of the accountability system. Nor does anyone know how this intermittent inclusion might affect judgments concerning alignment of assessments with standards.

Adding new tests into the system. Many states are in the process of phasing out their norm-referenced testing programs and have not yet finalized their standards-based programs. Roeber (2002) provides a useful way for thinking about this transition. Approaches range from adding items to available norm-referenced tests (NRTs) to align them with state academic content standards and grade-level expectations (GLEs) to creating new standards-based tests.

Another issue for states is adding new tests into an existing system (e.g., adding tests for grades that previously were not tested in the IASA grade-span approach). How can new tests be linked to the existing tests or system? Clearly, vertical alignment becomes an issue. Equating and readjusting targets may also be necessary, depending on results of a vertical alignment study. As discussed earlier, as states strive to differentiate standards from grade to grade, there is a danger that unless GLEs are vertically aligned, lower-grade assessments may contain more difficult items than higher grade assessments. At the very least, cut points will need to be "smoothed" (i.e., achievement-level cut points may need to be adjusted to offset the impact of misaligned GLEs from grade to grade).

Universally designed assessments. Access is a familiar topic for those concerned with assessing students with special needs. However, increasing attention to the concept of universal design has opened this discussion to all participants in the design and use of assessments. The concept of universal design was applied first to lesson design strategies for maximizing access for all learners to instruction. Increased accountability for the performance of all students has generated greater interest in how the concepts of universal design might be applied to large-scale assessment design. Universally designed assessments are developed from the beginning to be accessible and valid for the widest possible range of students (NCEO, 2003). The characteristics of universally designed assessments include an "inclusive assessment population; precisely defined constructs; accessible, non-biased items; amenable to accommodations; simple, clear, and intuitive instructions and procedures; maximum readability and comprehensibility; and maximum legibility" (Thompson, Johnstone, & Thurlow, 2002).

Substantial controversy surrounds the questions of appropriate methodology for universally designed assessments and their effectiveness. Some assessment experts contend that the skillful application of best practice in assessment and improvements in assessment design are already leading to universal access. Others contend that the principles of universal design may sound simple and logical but implementation is complex and costly. Further, retrofitting existing tests to increase access is proving even more costly than applying principles of universal design from the outset. In addition, differentiating universal design from methods, such as the use of "plain language," to meet the needs of specific subgroups is proving difficult.

Advocates of universal design emphasize the need to address the full range of students. They suggest this target population affects test conceptualization, test construction, test tryout procedures, item analysis procedures, and test revision. While little concrete evidence yet exists regarding effectiveness of universal design in increasing access to assessments and accuracy of results, the National Center for Educational Outcomes (NCEO) has been engaged in a project to examine the comparative impacts of traditionally and universally designed assessments. A recent NCEO report suggests that universally designed assessments are useful for increasing the validity of test results for students with disabilities and English language learners (Johnstone,

2003). However, with research on this topic still in its infancy, the extent to which the application of these principles impacts the constructs being assessed remains unclear. This is particularly true for plain language assessments, which some categorize as alternate assessments for English language learners and others categorize as simply another application of universal design. States will need to proceed cautiously as they determine whether and how to apply principles of universal test design.

Equity Issues and Special Cases

Grade-level expectations. Just as it seemed there was nearly agreement on definitions of vocabulary for talking about content and performance standards and their components, such as performance descriptors (e.g., Hansche et al., 1998), a new vocabulary has arisen to accommodate the requirements of NCLB. These emerging terms are not universally used or understood but, in essence, they concern the expectations for student achievement and performance at the end of every grade level. Under IASA, it was sufficient to benchmark standards and assessments to the end of a grade span, such as fourth grade, eighth grade, and twelfth grade.

Louisiana illustrates the approach taken by a number of states in several ways: (a) revision of the state content standards; (b) development of GLEs, keyed to the standards, for every grade in specified content areas, i.e., English language arts, mathematics, science, and social studies; (c) development and alignment of curricula with the GLEs; and (d) development and alignment of state assessments with the GLEs.

Other states have chosen to specify grade level standards keyed to graduation goals or exit standards. GLEs and grade-level standards serve approximately the same purpose. However, setting grade-level standards necessitates a full revision process, including completion of administrative procedures for the adoption of new standards. In either case, GLEs or grade-level standards provide an important link between content or exit standards and performance descriptors on the one hand, and curricula and instructional strategies, including classroom assessment, on the other. While GLEs or grade-level standards are helpful to teachers and curriculum developers in framing ongoing instruction, they present an equity issue relative to the timing of assessments. When state assessments for educational accountability purposes covered a grade span rather than a single grade, there was some legitimacy to administering a test of learning in grades 1 to 4 in the middle of the fourth grade, thereby allowing time for scoring and reporting by the end of the school year. However, to assess only fourth-grade learning in the middle of the fourth grade, especially in year-round schools, is problematic. States are grappling with the issue of whether to be so specific about the pacing of curricula that tests can accurately assess what is achieved by mid-year. This strategy could limit teacher flexibility to meet the needs of individual students relative to learning speed or sequence.

In states that do not have year-round schools, some propose setting cut scores that take into account the time of year at which the test is administered. This may help to account for the level of achievement students should have reached by that point in that grade level. For states that provide for multiple administrations, ED permits states to “bank” student scores of proficient

or higher on tests administered prior to the state's "first administration" of an assessment. The first administration is the first time an assessment is officially administered to measure a student's achievement of the state's academic content standards in the grade or subject for which the state expects the student to have achieved mastery of those standards. A score from an assessment administered to a student prior to this time would not be "banked" unless the student achieved a proficient or higher level. In these circumstances, issues of test security and validity must be addressed. A similar method might be used in year-round and other non-traditional situations. However, states have found issues of reporting and calculating participation rates to make the use of such procedures complicated and particularly challenging to existing databased systems.

Comparing students across schools. Equity, here, refers to the fairness of how schools are compared to one another based on student test results. This is particularly problematic if some schools are year-round and others are not, because there could be as many as 30 days difference in instruction. Some states handle the differences in number of days of instruction and year-round school issues by specifying the number of days per grade students should have completed by the time they are tested. Some states intend to administer fall tests that measure the previous grade-level standards, thus neutralizing this issue for all students. This procedure has the added appeal of mitigating time issues for data analysis and reporting results. However, it also creates a lag in the provision of choice, supplemental educational services, and other NCLB sanctions.

Students in alternate programs. The law is clear that annual accountability determinations must be made for all schools, including alternative schools. According to the ***Federal Register*** (December 2, 2002, p. 71744), ED has plans to issue guidance and examples for handling such cases. Special purpose schools of various descriptions fit this category. Students may attend these schools due to unique special education needs or for other special purposes. While any student who attends more than one school in a district during the school year is included only in the determination of district-level AYP, some of these students remain in alternate programs for extended periods or several years. Some states "backtrack" students to their home schools and report their scores for AYP purposes as part of that school's results. These states believe this procedure neutralizes the impact of special purpose schools that have high concentrations of special populations and has the added benefit of increasing the home school's interest in ensuring that the placement will maximize that student's achievement. This backtracking method is also used by many states to hold K-2 schools accountable. Scores are backtracked and used for AYP calculations for K-2 schools from the first grade tested, usually grade 3 or 4 at the receiving schools.

End-of-course tests. Some states (e.g., New York) use end-of-course tests to calculate AYP. This practice raises several issues. For example, is the course a "common" course required of all students? If so, it simplifies the calculation of AYP, but it may also encourage the use of a "low level" course that all students may be expected to take and, eventually, pass. Does the test represent a class or a course of study that may encompass more than one class? Some schools/states/districts divide Algebra I into a two-year course, for example, for some or all of their students. In cases where this option is provided only for some students, the intention is to

make time the variable for learning, thereby maximizing each student's opportunity to meet the standards.

States using end-of-course tests must ensure all students are assessed on the standards each must meet for graduation. For example, even though NCLB does not link passing state assessments to eligibility for graduation, if the state requires all students achieve Algebra I and geometry standards for graduation, all students are assessed for both. For AYP purposes, scores for students who must repeat courses for graduation count at the point the students would have been expected to complete that course. Subsequent scores of repeat course takers are not included, and scores for early course completers are banked for inclusion in AYP calculations.

Alternate assessment. Both IASA and NCLB provide for alternate assessments for a small number of students with significant physical or cognitive disabilities that would prevent them from meaningfully participating in the same assessments as other students, even with appropriate accommodations. States were required to establish guidelines for alternate assessments of students with disabilities (SWDs). Section 200.6(a)(2)(i) of the final regulations for standards and assessments under NCLB states that "the State's academic assessment system must provide for one or more alternate assessments for a student with disabilities as defined under section 602(3) of the IDEA who the student's IEP team determines cannot participate in all or part of the State assessments under paragraph (a)(1) of this section, even with appropriate accommodations." This suggests that more alternate assessments might be administered under NCLB than under IASA.

In March 2003, ED proposed a rule that would permit states to measure students with the most significant cognitive disabilities (not more than 1 percent of the total student enrollment) against alternate achievement standards aligned with the state's content standards and reflective of the highest learning standards possible for such students. The final amended regulations for special education under Title I were issued December 2003. These rules permit states to establish alternate achievement standards for students with the most significant cognitive disabilities. Alternate assessments aligned to the alternate achievement standards may be used to determine the proficiency of these students when calculating AYP. However, the regulations establish a 1 percent cap on the number of proficient and advanced scores that may count toward AYP, although no limit is placed on the number of students who may take the alternate assessment. Further, although this cap applies to AYP determinations for states and districts, parents must be notified of the students' actual performance on the alternate assessment (Coleman & Palmer, 2003).

Alternate assessments did not receive close scrutiny during peer reviews for IASA. Under NCLB, these assessments are likely to be a priority for review, with particular attention given to technical aspects of their design and implementation. Even though more than one alternate assessment might be allowable under NCLB, all would be required to meet high standards for technical quality, a consideration that has not been paramount in the design of many such assessments for most states. Further, alternate assessments that are fully aligned to state standards might be allowable if they can be proven to equal the level of the standards. However, the technical challenges of designing and implementing such assessments, documenting their technical quality, and verifying full alignment with the range, breadth, depth, level and degree of

cognitive complexity, and comprehensiveness of the standards present significant fiscal and human resource challenges for states.

Accommodations. While states continue to wrestle with issues of accommodation, the requirements remain essentially unchanged from IASA to NCLB. Accommodations are changes in the administration of an assessment, such as setting, scheduling, timing, presentation format, response mode, or others, including any combination of these. To yield valid results, accommodations must be ones that are also used in instruction, and they must not change the construct intended to be measured by the assessment or the meaning of the resulting scores (Redfield, 2002). Oregon has established a review board to ensure the validity of accommodations. Expert board members review unique requests to ensure access and validity.

Students in the "gap." In amended regulations (December 2003), ED clarified that states are permitted to develop alternate assessments that are aligned with the state's academic content standards for students with disabilities who are (1) unable to participate in the regular assessment with accommodations and (2) not significantly cognitively impaired. These alternate assessments must be designed to measure proficiency based on grade-level achievement standards. Proficient scores on alternate assessments aligned with grade-level standards are not subject to the 1 percent cap. However, these assessments must meet the same rigorous alignment and other technical requirements as the state's regular assessment.

One approach that states have used in the testing of "gap" students is out-of-level testing. In the first peer reviews of state accountability plans by ED under IASA, the practice of out-of-level testing was rejected. Later, ED accepted the use of out-of-level testing, providing the results were counted as "nonproficient." In a June 27, 2002, letter to the chief state school officers, Secretary Paige reversed all earlier ED decisions relative to out-of-level testing for 2002-03 only:

> . . . if a state permitted the use of ILAs [Instructional Level Assessments] during the 2002-2003 school year to measure the progress of other students with disabilities [other than those with the most severe cognitive disabilities for whom proposed regulations would permit the use of alternate assessments provided that the percentage of students held to related alternate standards does not exceed 1 percent of all students in the grade assessed] based on their IEPs, the State may hold schools and districts accountable for the achievement of these students against instructional-level standards rather than grade level standards. This policy only applies to assessments that were administered during the 2002-2003 school year, which will be used to make AYP determinations for the 2003-2004 school year. The 1.0 percent limit . . . does not apply to these test scores.

Under the December 2003 amended regulations, as interpreted by ED, states are permitted to use out-of-level assessment if the state meets the requirements for alternate achievement standards. Achievement standards for out-of-level assessments are clearly different from the achievement standards in the target grade, but out-of-level assessments that are

administered to students with the most significant cognitive disabilities and that meet the requirements of the regulation may be considered to be alternate assessments based on alternate achievement standards. Out-of-level tests used for students based on the regular standards, however, would require a totally separate set of achievement standards for each variant out-of-level use. For example, if an eighth-grade student is administered a fifth-grade standards-based test, a separate set of achievement standards would be needed to identify that student's proficiency level because the student is not a fifth-grade student taking a fifth-grade test. All out-of-level tests also will be subject to the 1 percent cap.

Conclusions. The issues described above require states' attention to the kinds of assistance they request from test developers, regardless of whether the state is adding grade-level tests into an existing system or transitioning from a NRT-based assessment system to a standards-based assessment system. The topics raised here will be discussed in greater detail in subsequent papers in this series. ED's Peer Review Guidance for standards and assessments under NCLB, recently released, should be useful in further clarifying these topics and related concerns. To clarify some of the questions that states need to address in transitioning to full compliance with NCLB requirements, the following Decision Tool is suggested.

Decision Tool

The following checklist may be considered in developing assessment RFPs, evaluating RFPs, working with assessment developers and vendors, and making related decisions. These questions are not intended to be all-inclusive but are important considerations in assessment development.

Questions to Ask of Potential Developers of Standards-Based Assessments

1. What experience do you have developing, validating, and implementing standards-based assessments? Who are your prior and current state clients? Which grades are included in your current test development contracts?

2. To what extent have you worked with states to develop academic content standards? Achievement standards? Performance descriptors? Grade-level expectations? Have you conducted external alignment studies? Which states?

3. What kinds of technical assistance have you provided or are you able to provide to states in building the capacity of LEAs and the SEA to meet the requirements of NCLB?

4. Describe the step-by-step process used to develop and validate assessments. Which alignment protocol do you use (e.g., Webb, Porter)?

5. Do you have enough items in your item banks to develop multiple test forms? What kinds of technical procedures have been applied to examine item characteristics and technical quality?

6. What has been your experience in high-stakes environments? For example, have your assessments and procedures withstood testing in courts of law? In which states have you and the state been successful in the courts?

7. What is the cost per student per test for an augmented NRT versus a new, standards-based assessment? This cost should include item development costs as well as testing materials, scoring, and reporting costs.

8. What is your experience in meeting demanding timelines? What is your typical turnaround time from the first day of testing until the final release of student-level results?

9. Explain your process for scoring open-ended items.

10. What is the turn-around time for the scoring and reporting of selected-response versus open-ended items/tests?

11. Can you provide reports per our specifications that meet NCLB reporting requirements and that are suitable for parents, teachers, and school report cards?

12. How have you addressed equity issues with regard to provision of accommodations to students with disabilities and limited English proficient students? How have you addressed these issues with regard to Section 504 students?

13. How have you applied concepts of universal design to assessments you have developed?

14. Have you developed alternate assessments? If so, in what areas? On what standards were they based? For which states have you developed alternate assessments and at which grades?

15. Can you help us manage the large databases associated with our testing program? What kind of experience in handling large databases do you have with your current state contracts? Which states? Which grades? How many students per grade?

16. Can you make data available in a manner useful for instructional purposes at all levels?

17. How do I know that you understand NCLB requirements? How does your company keep abreast of the requirements and guidance for NCLB?

18. Can you help us answer policymakers/board members' questions in terms that are clear and readily understandable?

Following are a series of steps to guide transitioning from an NRT-based assessment system to a standards-based assessment system.

Steps for Transitioning an NRT-Based Assessment System into a Standards-Based Assessment System

1. Determine if you wish to discontinue the administration of NRTs altogether at the state level. If not, it may be sufficient to your purposes to administer them to only a few grades or to test different subjects in different grades. If the purpose is to obtain a score on every student, perhaps because parents desire such, then every student at the designated grade and in the designated content area will need to be tested and assigned a score. If the purpose is to gauge the standing of the school or district relative to a national norm, then it may be possible to administer the NRT on a sampling basis. Scores from a NRT are not designed for making judgments about the proficiency level of a student. Proficiency is a standards-based phenomenon.

2. Whether or not you continue to administer NRTs, you will need to adapt or develop standards-based assessments for reading or language arts and mathematics in grades 3-8 and at least once in grades 10-12. By the 2007-08 school year, you will also need to develop standards-based assessments in science to be administered at least once during each of three grade spans: 3-5, 6-9, and 10-12.

3. Determine whether you will contract for the development of an augmented NRT or the development of new standards-based assessments built to specifically align with your standards and grade-level expectations. Ultimately, this decision may be made on the basis of cost. Other influencing factors would include whether you want to be able to use the data to make norm-based comparisons. For your state, what are the tradeoffs?

4. Develop, review, or revise state academic content standards, achievement standards, and grade- level expectations to serve as a solid foundation for the standards-based assessments.

5. Conduct alignment studies to ensure that the content standards, achievement standards, performance descriptors, grade-level expectations, test blueprints, and assessments are in horizontal and vertical alignment relative to depth of knowledge, range of knowledge, balanced representation, and source of challenge.

6. Pilot and field-test the new assessments. Adjust items and procedures as warranted. Establish the reliability and validity of the new assessments relative to the kinds of decisions they will be used to inform.

References

Carlson, D. (1996). *Adequate yearly progress provisions of Title I of the Improving America's Schools Act: Issues and strategies*. Washington, DC: Council of Chief State School Officers, State Collaborative on Assessment and Student Standards, Comprehensive Assessment Systems (CAS).

Coleman, A., & Palmer, S. (12 December 2003). *Summary and analysis of the U.S. Department of Education's final regulations on alternate achievement standards and alternate assessments under the No Child Left Behind Act of 2001*. Washington, DC: Council of Chief State School Officers.

Erpenbach, W. J., Forte Fast, E., & Potts, A. (2003). *Statewide educational accountability under NCLB*. Washington, DC: Council of Chief State School Officers, State Collaborative on Assessment and Student Standards, Accountability Systems and Reporting (ASR).

Hansche, L. N., Hambleton, R. K., Mills, C. N., Jaeger, R. M., & Redfield, D. L. (1998). *Handbook for the development of performance standards: Meeting the requirements of Title I*. Prepared for the U.S. Department of Education and the Council of Chief State School Officers, Washington, DC.

Improving America's Schools Act of 1994, Pub. L. 103-382. (1994). Retrieved from http://www.ed.gov/legislation/ESEA/toc.html

Johnstone, C. J. (2003). *Improving validity of large-scale tests: Universal design and student performance* (Technical Report 37). Minneapolis: University of Minnesota, National Center on Educational Outcomes. Retrieved from http://education.umn.edu/NCEO/OnlinePubs/Technical37.htm

Langenfeld, K. L., Thurlow, M. L., & Scott, D. L. (1996). *High stakes testing for students: Unanswered questions and implications for students with disabilities* (Synthesis Report No. 26). Minneapolis: University of Minnesota, National Center on Educational Outcomes. Retrieved from http://education.umn.edu/NCEO/OnlinePubs/Synthesis26.htm

Lissitz, R.W., & Huynh, H. (2003). Vertical equating for state assessments: Issues and solutions in determination of adequate yearly progress and school accountability. *Practical Assessment, Research & Evaluation, 8*(10). College Park, MD: PAREonline. Retrieved from http://PAREonline.net/getvn.asp?v=8&n=10

Marion, S., White, C., Carlson, D., Erpenbach, W. J., Rabinowitz, S., & Sheinker, J. (2002). *Making valid and reliable decisions in determining adequate yearly progress*. Washington, DC: Council of Chief State School Officers, ASR-CAS Joint Study Group on Adequate Yearly Progress. Available: http://www.ccsso.org/publications/details.cfm?PublicationID=57

National Center for Educational Outcomes. (2004). *Special topics: Universally design assessments: Frequently asked questions about universally designed assessments.*

Minneapolis: University of Minnesota, National Center on Educational Outcomes. Retrieved from http://www.education.umn.edu/nceo/TopicAreas/UnivDesign/UnivDesign_FAQ.htm

National Governor's Association. (2002). *NGA summary of the timeline requirements of No Child Left Behind*. Washington, DC: National Governor's Association (NGA). Retrieved from http://www.nga.org/center/divisions/1,1188,C_ISSUE_BRIEF^D_4767,00.html

No Child Left Behind Act of 2001, Pub. L. No. 107-110. (2002). Retrieved from www.ed.gov/legislation/ESEA02/indexhtml

Redfield, D. L. (2001). *Critical issues in large-scale assessment*. Washington, DC: Council of Chief State School Officers, State Collaborative on Assessment and Student Standards, Technical Guidelines for Performance Assessments (TGPA).

Roeber, E. D. (2002, November). *Policy brief: National forum on accountability*. Denver: Education Commission of the States. Available at www.ecs.org

Thompson, S. J., Johnstone, C. J., & Thurlow, M. L. (2002). *Universal design applied to large-scale assessments* (Synthesis Report 44). Minneapolis: University of Minnesota, National Center on Educational Outcomes. Available: http://education.umn.edu/NCEO/OnlinePubs/Synthesis44.html

U. S. Department of Education. (2002). *Information quality guidelines*. Retrieved from http://www.ed.gov/policy/gen/guid/infoquality.html?exp=0

U. S. Department of Education. (2003). *Title I regulations on alternate achievement standards: Questions and answers*. Retrieved from http://www.state.sd.us/deca/TA/basic/pdf/Q&A_on_Alternate_Achievement_Standards.pdf

U. S. Department of Education, Office of Elementary and Secondary Education. (2003, December 9). Title I-Improving the academic achievement of the disadvantaged: Final rules. *Federal Register* 68 (236). Washington, DC: Government Printing Office. Retrieved from http://www.ed.gov/legislation/FedRegister/finrule/2003-4/120903a.html

Webb, N. L. (2002). *Alignment study in language arts, mathematics, science, and social studies of state standards and assessment for four states*. Washington, DC: Council of Chief State School Officers.

Winter, P. C. (1996). *Implementing the adequate yearly progress provisions of Title I in the Improving America's Schools Act of 1994*. Washington, DC: Council of Chief State School Officers, State Collaborative on Assessment and Student Standards, Comprehensive Assessment Systems (CAS).

Wisconsin Department of Public Instruction. (2003). *Information for schools preliminarily identified for improvement*. Madison: Wisconsin Department of Public Instruction. Retrieved from http://www.dpi.state.wi.us/oea/doc/reconsid02.doc

APPENDIX A

Summary of Key Accountability Requirement Differences Between the 1994 and 2001 ESEA Reauthorizations[1]

	1994 ESEA Reauthorization (Improving America's Schools Act)	2001 ESEA Reauthorization (No Child Left Behind Act)
Transition Period	Almost one full school year with additional time to bring on line aligned standards, assessments, and accountability systems.	None—law was effective on enactment although standards and assessments beyond those previously required will be phased in gradually
Assessments	Reading or language arts and mathematics at least once annually in the three grade spans: 3-5, 6-9, and 10-12.	Same except that the assessments must be administered at least once annually in each grade, 3 through 8 by 2005-06 (and once within grades 10-12) with science administered at least once in each of the three grade spans by 2007-08.
Statewide Accountability Systems	Statewide system, using assessments administered to all students (not just those in Title I) to measure progress of schools and districts participating in Title I.	Single, statewide system required to measure progress of all schools and districts, not just those participating in Title I.[2]
Adequate Yearly Progress (AYP) Measures	States were required to establish AYP standards that could be limited to schools and districts receiving Title I funds. Identification of schools and districts based on the performance of *all* students, with no pre-determined annual growth rate or period of time for all students to master a state's academic content standards. Multiple measures of student performance could be applied in AYP determinations.	Each of at least nine subgroups of students must reach proficient or advanced achievement levels in reading or language arts and mathematics by 2013-14 (uniform progress is required beginning in 2002-03). AYP determinations are based solely on student achievement results in State assessments. At least 95% of the students in each subgroup must participate in the assessments and all must meet the state's performance target in another academic indicator as prescribed in the law.

[1] Source: Marion, S., White, C., Carlson, D. Erpenbach, W. J. Rabinowitz, S., & Sheinker, J. (2002*). Making valid and reliable decisions in determining adequate yearly progress*. Washington, DC: Council of Chief State School Officers, ASR-CAS Joint Study Group on Adequate Yearly Progress.

[2] On occasion, ED has made exception to these requirements on a state-by-state basis. See Erppenbach, Fort Fast, and Potts (2003).

	1994 ESEA Reauthorization (Improving America's Schools Act)	2001 ESEA Reauthorization (No Child Left Behind Act)
Rewards and Sanctions	*Rewards*: States were to identify especially successful schools and distinguished educators and were authorized to use Title I funds to provide additional support. *Sanctions*: Many possibilities identified but most could not be taken until standards and assessments were fully implemented.	*Rewards*: States must identify rewards and may use Title I funds in support of the rewards. *Sanctions*: A set of progressive sanctions required to be applied to low-performing schools and districts receiving Title I funds. Most sanctions are automatic although districts and states have some discretion regarding the extent of the number and scope of sanctions related to corrective actions that are applied under the law.
English Language Acquisition	Acquisition of English language proficiency not required.	States must set annual measurable objectives for increasing English language proficiency by limited English proficient students, and districts must annually assess same.
National Assessment of Educational Progress	State, district, and school participation not required.	State participation required as well as district and school participation if district receives Title I funds.

APPENDIX B
NCLB Implementation Timeline [3]

Standards

Key Dates	NCLB Provision
Immediately	Standards and grade-level expectations must be developed for reading or language arts and mathematics in grades 3-8.
May 2003	States must submit evidence of reading or language arts and mathematics academic content standards and grade-level student academic achievement (performance) standards.
2005-06	States must implement reading or language arts and mathematics 3-8 grade-level student achievement standards.
2005-06	Science academic content standards must be developed for elementary, middle, and high school.
May 2006	States must submit evidence of science academic content standards and student academic achievement standards.
December 2006	States must submit evidence of implementing reading or language arts and mathematics student academic achievement standards.

Assessment/Testing

Key Dates	NCLB Provision
2002-03	States must administer assessments in reading or language arts and mathematics at least once in grade spans 3-5, 6-9, and 10-12.
2005-06	States must administer assessments every year in grades 3-8 (and at least once in grades 10-12) in reading or language arts and mathematics.
December 2006	States must submit evidence of implementing reading or language arts and mathematics assessments in grades 3-8 as well as once in grades 10-12.
2007-08	States must administer assessments in science at least once in grade spans 3-5, 6-9, and 10-12.
December 2008	States must submit evidence of implementing science assessments.

[3] Adapted from National Governor's Association. (2002). *NGA summary of the timeline requirements of No Child Left Behind*. Washington, DC: National Governor's Association (NGA). Retrieved from http://www.nga.org/center/divisions/1,1188,C_ISSUE_BRIEF^D_4767,00.html

NAEP Assessment

Key Dates	NCLB Provision
Starting in 2002-03, continuing every other year	The U.S. Department of Education will pay for a state to participate in the NAEP reading and mathematics assessment for 4th- and 8th-graders every other year.

LEP Student Assessment

Key Dates	NCLB Provision
Starting in 2002-03	States must ensure that districts administer annual assessments of English language proficiency to LEP students.

Report Cards

Key Dates	NCLB Provision
Beginning of 2002-03	States must "promptly" provide the results of assessments no later than before the beginning of the next school year to LEAs, schools, and teachers in a manner that is clear and easy to understand.
Beginning of 2002-03	Districts disseminate annual local report cards.

APPENDIX C
Within Year Requirements[4]

The Council of Chief State School Officers developed a planning matrix for use by states. This planning matrix sets forth descriptions of requirements and dates requirements need to be completed related to:

- Standards and Assessments
- Accountability/AYP
- School Improvement
- Data and Reporting
- Teacher Quality
- Special Populations (ELL, LEP)
- State (and USED) Monitoring
- Leadership Quality

The timeline provides within-year requirements delineated in particular detail for 2004. To access the document see:

http://www.ccsso.org/content/pdfs/TimelineNCLB.pdf

[4] Provided by the Council of Chief State School Officers (http://www.ccsso.org/content/pdfs/TimelineNCLB.pdf)